GRAFT: Decoupling Ranking and Calibration for Survival Analysis

Mohammad Ashhad MOHAMMAD.ASHHAD@KAUST.EDU.SA
CEMSE, KAUST, KSA

Robert Hoehndorf ROBERT.HOEHNDORF@KAUST.EDU.SA
CEMSE, KAUST, KSA

Ricardo Henao RICARDO.HENAO@DUKE.EDU
Dept. of Bioinformatics & Biostatistics, Duke University, USA

Abstract

Survival analysis is complicated by censored data, high-dimensional features, and non-linear interactions. Classical models offer interpretability and superior calibration but are restricted to linear or predefined functional forms, while deep learning models are flexible and achieve strong discriminative performance, but tend to produce poorly calibrated survival estimates. To address this trade-off, we propose GRAFT (Gated Residual Accelerated Failure Time), a novel AFT model that decouples prognostic ranking from survival calibration. GRAFT's hybrid architecture combines a linear AFT model with a non-linear residual neural network, and it also integrates stochastic gates for automatic feature selection. The model is trained by optimizing a differentiable, C-index-aligned ranking loss using stochastic conditional imputation from local Kaplan-Meier estimators, while calibrated survival estimates are obtained through simple post-training calibration. In public benchmarks, GRAFT outperforms baselines in discrimination and calibration, while remaining robust and sparse in high-noise settings.

1. Introduction

Survival analysis, or time-to-event modeling, is a foundational branch of statistics focused on analyzing the expected duration of time until one or more events occur (Clark et al., 2003). This methodology is indispensable in fields where the "when" is as critical as the "if". In clinical medicine, it is used to predict the prognosis of the patient (He, 2023), assess the efficacy of new treatments by modeling the time to relapse (Kassani et al., 2015), and identify risk factors for mortality (Abbasi et al., 2024). In engineering, it serves as a reliability analysis by predicting the time-to-failure of mechanical components (Meeker et al., 2014). In finance, it is used to model the time-to-default in credit risk scoring (Dirick et al., 2017).

The principal challenge that distinguishes survival analysis from standard regression is the presence of censoring. A subject is right-censored if the event of interest has not occurred by the end of the study period, or if the subject is lost to follow-up. Standard regression techniques, such as Ordinary Least Squares (OLS), would yield biased estimates if applied to these types of partially observed data. Consequently, a specialized set of models has been developed to handle censoring appropriately (Clark et al., 2003).

For decades, the field has been dominated by two classical approaches. The first is the non-parametric Kaplan-Meier (KM) estimator (Kaplan and Meier, 1958), which provides a simple estimate of the (marginal) survival function but cannot incorporate features or covariates. The second and most prominent is the Cox Proportional Hazards (CoxPH) model (Cox, 1972). As a semi-parametric model, CoxPH is highly valued for its calibration and its interpretable log-linear relationship between features and a *hazard ratio*. However, its validity rests on the strict Proportional Hazards (PH) assumption that the effect of a covariate on the hazard is constant over time. This assumption is frequently violated in real-world data. Parametric Accelerated Failure Time (AFT) models (Saikia and Barman, 2017) offer an alternative by assuming a direct, linear relationship between covariates and the logarithm of survival time, but, in turn, they are limited by rigid distributional assumptions (*e.g.*, Weibull or Log-Normal).

The advent of large-scale, complex and noisy datasets such as high-dimensional genomic data or granular Electronic Health Records (EHRs) has further exposed the limitations of these classical models. The high-dimensionality of modern datasets introduces a critical need for robust feature selection. Although methods like LASSO (L1 regularization) can be applied to CoxPH models as a feature selection mechanism (Tibshirani, 1997), they exhibit selection instability in the presence of correlated predictors (Zou and Hastie, 2005; Zhao and Yu, 2006). Beyond feature selection, the non-linear and complex interactions present in modern datasets have spurred the development of machine learning and deep learning approaches, such as Random Survival Forests (RSF, Ishwaran et al., 2008) and various deep neural networks such as DeepSurv (Katzman et al., 2018). Although these models often achieve superior discriminative performance by capturing complex patterns, they tend to produce poorly calibrated survival estimates. Deep learning models, in particular, are also prone to overfitting and can be sensitive to irrelevant or "noisy" features.

To bridge the gap between the calibration strength of classical models and the discriminative power of deep learning, we propose the **Gated Residual AFT (GRAFT)** model. A key design principle of GRAFT is the explicit decoupling of prognostic ranking from survival calibration. Our model is a novel AFT architecture designed to be flexible, robust, and sparse simultaneously. It introduces a hybrid framework that learns both linear and non-linear feature effects, while employing a stochastic gating mechanism to perform automatic feature selection during training. The primary contributions of this work are threefold.

1. **Decoupling of Ranking and Calibration:** We propose an explicit separation between prognostic ranking and survival calibration. GRAFT is trained to optimize discrimination via a differentiable ranking objective, while calibrated survival estimates are obtained through a simple post-training calibration step.
2. **A Hybrid AFT Architecture:** We combine a linear AFT model with a non-linear residual Multi-Layer Perceptron (MLP), allowing the model to capture both simple linear effects and complex non-linear interactions. The model is trained using stochastic conditional imputation from local Kaplan-Meier estimators to handle censored observations, and optimized by directly minimizing a differentiable soft-ranking loss via Monte Carlo averaging, in alignment with the Concordance Index (C-index, Harrell et al., 1982).
3. **Integrated Stochastic Feature Selection:** We incorporate Stochastic Gates (STG) (Yamada et al., 2020), a differentiable feature selection mechanism based on continuous

relaxation of Bernoulli variables, directly into the model architecture. This enables the model to automatically identify relevant features and suppress noisy or redundant inputs during training, providing robustness in high-dimensional settings.

We validate our model on six public benchmark datasets (GBSG, METABRIC, SUPPORT, NWTCO, FLCHAIN, and AIDS), demonstrating that GRAFT consistently outperforms both classical and modern baselines in both ranking and calibration. We further conduct experiments with injected noise features, proving that our model's stochastic gating mechanism successfully identifies and discards irrelevant information, a critical capability that traditional models lack.

Generalizable Insights about Machine Learning in the Context of Healthcare

- **Decoupling ranking and calibration leads to improvements in survival analysis.** Deep learning survival models typically couple discrimination and calibration into a single objective, forcing a trade-off between the two. Our experiments demonstrate that explicitly separating these objectives, optimizing ranking during training, and post-hoc calibrating survival estimates yield models that are simultaneously better at discrimination and calibration, both of which are critical for clinical decision-making.
- **Integrated feature selection is critical for robustness in healthcare scenarios.** Real-world clinical datasets are frequently high-dimensional, containing irrelevant measurements, noisy biomarkers, and spurious correlations. Our experiments demonstrate that models without explicit feature selection mechanisms degrade significantly under such conditions, while GRAFT's stochastic gating consistently identifies and suppresses uninformative features.

2. Related Work

The methodologies for survival analysis can be broadly categorized into classical statistical models and modern machine learning approaches. Our proposed model, GRAFT, draws inspiration from both traditions.

Classical Survival Models For decades, the field of survival analysis has been dominated by a few robust and interpretable models. The most foundational of these is the non-parametric Kaplan-Meier (KM) estimator, which provides a population-level estimate of the survival function $S(t)$ from time-to-event data (Kaplan and Meier, 1958). Although essential for summarizing survival data, the KM estimator cannot incorporate individual-specific covariates to provide personalized predictions. To address this limitation, the Cox Proportional Hazards (CoxPH) model was introduced (Cox, 1972). The CoxPH model is a semi-parametric method that estimates the effect of covariates x on the hazard rate, $\lambda(t|x)$, at time t without making assumptions about the underlying shape (or distribution) of the baseline hazard, $\lambda_0(t)$:

$$\lambda(t|x) = \lambda_0(t) \exp(x^T \beta)\,, \tag{1}$$

The output of the CoxPH model, the hazard ratio $\exp(\beta)$, is highly interpretable and has made it the *de facto* standard in clinical research. However, it relies on the strict Proportional

Hazards (PH) assumption that the effect of covariates is constant over time, which is often violated in practice (Clark et al., 2003).

In contrast, parametric models such as the Weibull model, which can be formulated as an Accelerated Failure Time (AFT) model, constitute an alternative (Kalbfleisch and Prentice, 2002). AFT models assume that covariates act to rescale time directly, such as by accelerating or decelerating the time-to-event. Although fully parametric models can be more efficient if the underlying distribution is correctly specified, their rigidity is a significant limitation. The proposed GRAFT model is built upon the flexible AFT framework, but, as we show, it circumvents the need for strict distributional assumptions.

Deep Learning for Survival Analysis With the rise of high-dimensional data, traditional models that assume linearity (like CoxPH and AFT) have proven insufficient for capturing complex, non-linear patterns. This has led to the development of deep learning models for survival analysis. Although these models often achieve strong discriminative performance, a consistent finding in benchmark studies is that their calibration tends to be poor compared to classical approaches (Burk et al., 2024).

One of the most prominent early-proposed models is DeepSurv (Katzman et al., 2018). DeepSurv directly adapts the CoxPH model to a deep learning context by replacing the linear term $x^T\beta$ in (1) with a deep feed-forward neural network $h_\theta(x)$, and optimizes a modified Cox partial log-likelihood loss. While effective at modeling non-linear covariate effects, DeepSurv still inherits the limitations of the proportional hazards assumption.

Other approaches have sought to discretize time and reformulate the problem, enabling more flexible hazard modeling. DeepHit (Lee et al., 2018) is a landmark model in this area. The model discretizes the time horizon into a set of bins and learns the joint distribution of survival time and event type, making it a natural fit for competing risks. By optimizing a loss function that combines a log-likelihood for the observed event time and a rank-based loss for censored subjects, DeepHit can estimate the full survival probability curve without being restricted by the PH assumption. DeepHit combines a log-likelihood term for calibration and a ranking loss for discrimination into a single objective, controlled by a scalar trade-off parameter, coupling the two objectives.

GRAFT addresses the key limitations of existing deep learning approaches through two design choices. First, it explicitly decouples discrimination from calibration: the model is trained to optimize ranking, while calibrated survival estimates are obtained through a dedicated post-training step. Second, it integrates stochastic gates directly into the architecture for automatic feature selection, providing robustness to noise and overfitting in high-dimensional settings. Together, these choices position GRAFT as a model that combines the discriminative power of deep learning with the calibration reliability of classical approaches.

3. The Gated Residual AFT Model

The proposed model, Gated Residual AFT (GRAFT), is a survival model that integrates local nonparametric imputation, stochastic feature selection, and a non-linear residual MLP. The model is trained via minibatch stochastic gradient descent by directly optimizing a differentiable approximation of the rank-correlation between event times and model predictions using Monte Carlo averaging as explained in the following.

3.1. Problem Formulation

We consider a dataset $D = \{(x_i, t_i, \delta_i)\}_{i=1}^N$, where N is the number of subjects. For each subject i, $x_i \in \mathbb{R}^p$ is a vector of p baseline features, $t_i > 0$ is the observed event time and $\delta_i \in \{0, 1\}$ is the event indicator, with $\delta_i = 1$ denoting an observed event and $\delta_i = 0$ denoting right-censoring, *i.e.*, $t_i = \min(y_i, c_i)$ and $\delta_i = \mathbb{I}(t_i < c_i)$ given the true event time and the censoring time y_i and c_i, respectively, of which only one is observed as t_i; and $\mathbb{I}(\cdot)$ is the indicator function.

The classical Accelerated Failure Time (AFT) model (Kalbfleisch and Prentice, 2002) provides a linear foundation for survival analysis by letting

$$\log(T_i) = x_i^T \beta + \mu + \epsilon_i, \tag{2}$$

where $\beta \in \mathbb{R}^p$ is the vector of coefficients, μ is an intercept term, and ϵ_i is an error term. The challenge is that $\log(T_i)$ is unknown for subjects with censoring ($\delta_i = 0$), which AFT models circumvent by specifying the distribution of ϵ_i in the location-scale family so the loss for censored observations can be written in closed form. Moreover, standard AFT models assume a linear relationship between features and log-survival event time, which may be insufficient for capturing complex, non-linear interactions in high-dimensional survival data.

3.2. GRAFT Model Overview

Instead of directly modeling $\log(T_i)$ as in (2), GRAFT learns a prognostic score $s_i \in \mathbb{R}$ that is monotonic with the log-survival time. Specifically,

$$\tilde{x}_i = g_{\eta,i} \odot x_i \tag{3}$$

$$\phi_i = \tilde{x}_i + f_\theta(\tilde{x}_i) \tag{4}$$

$$s_i = \beta^T \phi_i + \mu, \tag{5}$$

where $\odot$ denotes element-wise product and $g \in (0, 1)^p$ in (3) is a p-dimensional vector with values sampled from a stochastic gate parameterized by $\eta \in \mathbb{R}^p$ used here to mask the input features x_i. During training, the gate values are sampled from a binary distribution using the reparameterization trick. Note that since the gate vector parameter η is shared between all subjects, it enables feature selection at the population level. This gating mechanism is detailed in Section 3.4.

The gated features in (3) are then fed through a non-linear (residual) MLP $f_\theta : \mathbb{R}^p \to \mathbb{R}^p$ with parameters θ. For simplicity, in our experiments we use a fully-connected network with two layers, tanh activation function and d_h hidden units; however, deeper architectures or alternative parameterizations can also be considered.

The score s_i is obtained in (5) as a linear combination with parameters $\beta \in \mathbb{R}^p$ and $\mu \in \mathbb{R}$ of the gated features $\tilde{x}_i$ and the residual MLP $f_\theta(\tilde{x}_i)$, thus the representation ϕ_i in (4) captures both the direct effect of selected features and their non-linear interactions learned by the residual network.

The complete set of model parameters is $\Psi = \{\eta, \theta, \beta, \mu\}$, denotes the parameters of the stochastic gate described in Section 3.4. All parameters are learned jointly during training. In the following, we detail the stochastic imputation procedure for handling censored observations in Section 3.3, the stochastic gate mechanism in Section 3.4, the training objective in Section 3.5, and post-training survival curve estimation in Section 3.6).

3.3. Stochastic Conditional Imputation via Local KM

To train our AFT model on censored data, we require target values representing the underlying survival times. Instead of assuming a parametric distribution for ϵ_i in (2) and using its conditional density function to model censored data as in standard AFT modeling (see Kalbfleisch and Prentice (2002) for details), or using an imputation mechanism based on point estimates, we employ a *stochastic conditional imputation* approach based on local and nonparametric survival estimation (Dabrowska, 1987). This is done to preserve the distributional uncertainty inherent in censored data by letting the model train with plausible event-time realizations rather than a single imputed value. However, we still make the standard assumption in survival analysis that given the input features X, the true event time Y and the censoring time C are independent, *i.e.*, $Y \perp\!\!\!\perp C \mid X$.

Local KM Estimation. For each censored observation i, *i.e.*, with $\delta_i = 0$, we construct a local neighborhood $\mathcal{V}(x_i) \subseteq \{1, \ldots, N\}$ in the input feature space (using the Euclidean distance with standardized features) by expanding outward from x_i until k neighbors with observed events ($\delta_j = 1$) are found; all neighbors encountered during this expansion, including censored ones, are retained in $\mathcal{V}(x_i)$, so $|\mathcal{V}(x_i)| \geq k$. We set $k = 10$.

Using only the subjects in $\mathcal{V}(x_i)$, we fit a KM survival estimator (Kaplan and Meier, 1958), denoted as $\hat{S}(t \mid x_i)$ to approximate the conditional survival function $S(t \mid x_i)$. Since we know that the subject i has survived beyond their censoring time t_i, we condition the survival distribution on this information:

$$\hat{S}(t \mid x_i, T_i > t_i) = \frac{\hat{S}(t \mid x_i)}{\hat{S}(t_i \mid x_i)}, \quad t > t_i. \tag{6}$$

This conditional survival function is a nonparametric estimate of the distribution of T_i given covariates x_i and survival beyond t_i. We use such a k-nearest-neighbor-based estimator for $S(t \mid x_i)$ because it is an instance of the Beran conditional product-limit estimator, which is consistent under conditional independent censoring (Beran, 1981; Dabrowska, 1989, 1987).

From a practical perspective, it is worth noting that the neighborhoods and corresponding local KM curves are computed once using the full training set prior to optimization, defining fixed conditional imputation distributions that are used throughout training.

Stochastic Sampling. Rather than summarizing the conditional survival distribution $S(t \mid x_i)$ to a single point estimate, *e.g.*, its expectation, we follow the principle of multiple imputation (Rubin, 1987), and draw samples directly from the conditional distribution. For uncensored subjects ($\delta_i = 1$), we simply use the observed log-time $Y_i^* = \log(t_i)$. For censored subjects ($\delta_i = 0$), we sample imputed event times using

$$T_i^* \sim \hat{F}(\cdot \mid x_i, T_i > t_i), \tag{7}$$

where $\hat{F}(t \mid x_i, T_i > t_i) = 1 - \hat{S}(t \mid x_i, T_i > t_i)$ is the conditional cumulative distribution function and then set $Y_i^* = \log(T_i^*)$. In practice, this is done by inverse-transform sampling (Devroye, 2006). During training, multiple independent imputation values ($M = 5$) are generated for censored subjects within each minibatch, allowing the learning objective to account for the uncertainty of imputation through Monte Carlo averaging, as detailed in Section 3.5.

3.4. Feature Selection with Stochastic Gates

To achieve sparsity and identify relevant features in noisy or high-dimensional settings, we integrate Stochastic Gates (STG) (Yamada et al., 2020), directly into the model architecture. STG uses a continuous relaxation of Bernoulli random variables based on Gaussian distributions to learn a probabilistic binary mask over the input features.

For the j-th feature index (where $j \in \{1, \ldots, p\}$), we follow the generating process

$$\begin{aligned} \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \\ g_{\eta_j,i} &= \max\left(0, \min\left(1, (\eta_j + \epsilon_i)\right)\right), \end{aligned} \tag{8}$$

where $\mathcal{N}(0, \sigma^2)$ is a Gaussian distribution with variance σ^2, η_j is an element of $\eta \in \mathbb{R}^p$, which denotes the learnable mean gate parameter, and $g_{\eta_j,i} \in (0, 1)$ is a gate value sample for component j of x_i. These are shared by all data points in a minibatch to ensure that all subjects are evaluated with the same feature mask, thereby learning population-level feature importance rather than subject-specific patterns. The mechanism in (8) uses the reparameterization trick (Miller et al., 2017), which allows gradient-based optimization of η.

During inference (*i.e.*, at test time) and for convenience, the stochastic sampling in (8) is replaced by the (deterministic) expected value of the gate, *i.e.*,

$$g_{\eta_j} = \max\left(0, \min\left(1, (\eta_j)\right)\right), \tag{9}$$

We use deterministic gates during inference to eliminate sampling variance, provide reproducible predictions, and reduce computational overhead due to prediction averaging (Yamada et al., 2020); which in some cases may lead to more robust results.

To encourage sparsity, we add a differentiable approximation of the ℓ_0 norm to the loss function, which penalizes the expected number of active (non-zero) gates:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{L0}} \sum_{j=1}^{p} \mathbb{E}[g_j > 0] = \lambda_{\text{L0}} \sum_{j=1}^{p} \Phi\left(\frac{\eta_j}{\sigma}\right), \tag{10}$$

where $\lambda_{\text{L0}} > 0$ is a hyperparameter controlling the strength of the sparsity penalty (we set $\lambda_{\text{L0}} = 0.01$), and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Gaussian distribution. The expectation $\mathbb{E}[g_j > 0] = P(\eta_j + \epsilon_j > 0) = \Phi(\eta_j/\sigma)$ follows from the properties of the Gaussian distribution (Yamada et al., 2020).

3.5. Training Objective: Differentiable Ranking Loss

We train GRAFT by optimizing a ranking-based objective that aligns with the C-index (Harrell et al., 1982), commonly used as a evaluation metric in survival analysis (with discussed below). Our loss function combines a differentiable approximation of Spearman's rank correlation with the STG penalty in (10) and standard weight regularization.

Differentiable Ranking Loss. We optimize the negative Spearman's rank correlation between the predicted scores and target log-times for each minibatch of size m, with censored events imputed as described in Section 3.3, to encourage predictions to match the order of the target event times. Since the ranking operator needed by the Spearman's rank correlation is non-differentiable, we use the differentiable soft-ranking operator of Blondel et al. (2020),

which provides a continuous relaxation to the rank function while preserving the ordering of the input scores. This operator is computed through isotonic regression using the Pool Adjacent Violators Algorithm (PAVA) (Best et al., 2000).

For a minibatch $\mathcal{B} = \{i_1, \ldots, i_m\}$ with predicted scores $s_\mathcal{B} \in \mathbb{R}^m$ and imputed targets $Y^*_\mathcal{B} \in \mathbb{R}^m$, the ranking loss is

$$\mathcal{L}_{\text{rank}}(s_\mathcal{B}, Y^*_\mathcal{B}) = -\frac{\text{cov}(\hat{R}_\tau(s_\mathcal{B}), R(Y^*_\mathcal{B}))}{\sigma_{\hat{R}_\tau(s_\mathcal{B})} \cdot \sigma_{R(Y^*_\mathcal{B})}}, \tag{11}$$

where $\hat{R}_\tau(s_\mathcal{B}) \in \mathbb{R}^m$ and $R(Y^*_\mathcal{B}) \in \mathbb{R}^m$ denote the soft-ranking vector of predicted scores and the (hard) ranking vector of imputed targets within the minibatch, respectively, $\text{cov}(\cdot, \cdot)$ and $\sigma(\cdot)$ denote covariance and standard deviation, respectively, and $\tau > 0$ in $\hat{R}_\tau(\cdot)$ controls the degree of smoothing of the relaxed soft rank function (we use $\tau = 0.1$). We present a sensitivity analysis to the parameter τ in Appendix C.5.

Monte Carlo Averaging over Imputations. To account for the uncertainty in imputed values for censored observations, we combine Monte Carlo averaging with minibatch stochastic gradient descent. For each minibatch $\mathcal{B}$, we generate $M = 5$ independent imputation values by sampling from the pre-fitted conditional KM distributions (as described in Section 3.3).

Specifically, for Monte Carlo draw $k \in \{1, \ldots, M\}$, we construct an imputed target vector $Y^{*(k)}_\mathcal{B} \in \mathbb{R}^m$. For uncensored subjects ($\delta_i = 1$), we set $Y^{*(k)}_i = \log(t_i)$ (fixed across all k). For censored subjects ($\delta_i = 0$), we sample $T^{*(k)}_i \sim \widehat{F}_i(\cdot \mid x_i, T_i > t_i)$ (re-sampled for each k as in (7)) and set $Y^{*(k)}_i = \log(T^{*(k)}_i)$. The ranking loss in (11) then becomes the average over M sets of imputed target events

$$\bar{\mathcal{L}}_{\text{rank}}(\mathcal{B}) = \frac{1}{M} \sum_{k=1}^{M} \mathcal{L}_{\text{rank}}(s_\mathcal{B}, Y^{*(k)}_\mathcal{B}). \tag{12}$$

This Monte Carlo averaging ensures that the model optimizes ranking consistency in expectation over the imputation distribution, rather than overfitting to any single imputed realization.

Complete Training Objective. The final loss function for each minibatch combines the Monte Carlo-averaged ranking loss from (12), the STG penalty from (10), and standard L_2 weight regularization on all model parameters $\Psi = \{\eta, \theta, \beta, \mu\}$,

$$\mathcal{L}_{\text{total}}(\mathcal{B}) = \bar{\mathcal{L}}_{\text{rank}}(\mathcal{B}) + \mathcal{L}_{\text{reg}} + \alpha_{\text{L2}} \|\Psi\|_2^2, \tag{13}$$

where $\alpha_{\text{L2}} > 0$ is the regularization hyperparameter (we use $\alpha_{\text{L2}} = 10^{-4}$). We minimize $\mathcal{L}_{\text{total}}$ using the Adam optimizer (Kingma and Ba, 2014) via minibatch stochastic gradient descent with a learning rate of 10^{-3} to minimize the expected soft-Spearman ranking loss, averaged over Monte Carlo draws from the fixed conditional imputation distributions.

Under the assumption of conditional independent censoring, *i.e.*, $Y \perp\!\!\!\perp C \mid X$, and standard regularity conditions for local KM estimators (see Dabrowska (1989) for details), the estimated imputation distributions $\widehat{F}(\cdot \mid x_i, T_i > t_i)$ in (7) are consistent estimators of the true conditional event-time law $T_i \mid x_i, T_i > t_i$. Consequently, since the C-index (Harrell

et al., 1982) quantifies the pairwise ranking performance, optimizing a differentiable rank-based surrogate constitutes a tractable objective to improve concordance under censoring, *i.e.*, improving rank correlation will also improve the C-index as long as $\widehat{F}(\cdot \mid x_i, T_i > t_i) \to F(T_i \mid x_i, T_i > t_i)$ (in probability).

3.6. Survival Function Estimation

The GRAFT model, as described in Section 3.2, produces a prognostic score $s_i \in \mathbb{R}$ that is monotonic with log-survival times and optimized for ranking, thus for concordance as discussed above. However, it does not directly produce the survival function $\hat{S}(t \mid x) = \mathbb{P}(T > t \mid x)$. To obtain survival function estimates from the learned GRAFT scores, we adopt a post-training calibration procedure using the popular CoxPH model (Cox, 1972), which we specify as follows.

$$\lambda(t \mid s) = \lambda_0(t) \exp(\beta_{\text{cox}} \cdot s), \tag{14}$$

where $\lambda_0(t)$ denotes the baseline hazard function, $\beta_{\text{cox}} \in \mathbb{R}$ is the regression coefficient associated with the GRAFT score obtained using (5). The CoxPH model is fit by maximizing the partial likelihood, yielding an estimate of β_{cox} along with the standard Breslow estimator for the baseline cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$ (Lin, 2007).

For a test subject with input features x_{te}, we first compute the corresponding GRAFT score s_{te} using (5). The estimated survival function is then given by

$$\hat{S}(t \mid x_{te}) = \exp\left(-\Lambda_0(t) \cdot \exp\left(\beta_{\text{cox}} \cdot s_{te}\right)\right) \tag{15}$$

where $\Lambda_0(t)$ is evaluated at arbitrary time points through linear interpolation of the Breslow estimator obtained from the training data. This procedure leverages the GRAFT scores as a learned feature representation and uses CoxPH as a post-processing layer to convert these scores into time-dependent survival probability estimates. Importantly, this step does not modify the GRAFT parameters Ψ or its discrimination ability (C-index), and only requires fitting a one-dimensional CoxPH model, which incurs minimal additional computation, but guarantees calibrated predictions as long as $\Lambda_0(t)$ holds for x_{te} (Lin, 2007). This is not a strong assumption because one usually assumes that $\{(x_{te}, t_{te})\}$ and $\mathcal{D}$ (the training dataset), are i.i.d., however, the assumption that the baseline hazard function $\lambda_0(t)$ is not dependent on covariates x may not hold in practice, which can indeed affect calibration guarantees. We empirically investigate whether this assumption limits the calibration performance by comparing the CoxPH-based approach in (14) with non-parametric isotonic regression calibration (Tibshirani et al., 2011), in Appendix C.3.

4. Experiments

All experiments were conducted on a system with an AMD Ryzen Threadripper PRO 5975WX CPU and NVIDIA RTX 5000 Ada GPU, with deep learning models trained on the GPU. Code for reproducing all experiments is available at: https://github.com/ashhadm/GRAFT.

4.1. Experimental Setup

Datasets: We evaluate GRAFT on six public survival analysis benchmarks spanning diverse censoring rates and application domains. The GBSG (German Breast Cancer Study Group) dataset (Schumacher et al., 1994) contains 2,232 breast cancer patients with 43% censoring. The METABRIC dataset (Curtis et al., 2012) provides genomic data for 1,904 breast cancer patients with 42% censoring. The SUPPORT dataset (Knaus et al., 1995) includes 8,873 critically ill hospitalized patients with 32% censoring. The NWTCO dataset (Breslow and Chatterjee, 1999) comprises 4,028 pediatric kidney tumor patients with 86% censoring. The FLCHAIN dataset (Dispenzieri et al., 2012) contains 7,874 subjects from a serum free light chain study with 72% censoring. Finally, the AIDS dataset (Hammer et al., 1997) includes 1,151 HIV/AIDS patients with 92% censoring. This collection spans low-censoring (32-43%) and high-censoring (72-92%) scenarios. See Appendix C.6 for results with TCGA-BRCA (The Cancer Genome Atlas Network, 2012), a high dimensional genomic dataset.

Baselines: We compare GRAFT with five representative survival models. Cox Proportional Hazards (CoxPH) (Cox, 1972) serves as the standard semi-parametric baseline. Weibull AFT (Wei, 1992) represents a fully parametric accelerated failure time model. DeepHit (Lee et al., 2018) is a deep learning model that discretizes time and learns joint distributions without assuming proportional hazards. DeepSurv (Katzman et al., 2018) adapts the Cox model to deep learning by using neural networks to learn non-linear covariate effects. The Random Survival Forest (RSF) (Ishwaran et al., 2008) provides an ensemble tree-based method. The hyperparameters for all models are listed in Appendix B. To validate that our ranking loss serves as a reliable training objective for concordance, we also record the training loss and test C-index at every epoch across all runs of GRAFT.

Metrics and Evaluation: We evaluate both ranking and calibration performance using two complementary metrics. Harrell's C-index (Harrell et al., 1982) measures the probability that, for a random pair of subjects where one experiences the event before the other, the model correctly ranks their risk scores. Higher values indicate better discrimination, with the metric ranging from 0.5 (random) to 1.0 (perfect). The Integrated Brier Score (IBS) (Graf et al., 1999) measures the mean squared error between predicted survival probabilities and observed outcomes over time, weighted by the inverse probability of censoring (IPCW). Lower values indicate better calibration, with the metric ranging from 0 (perfect) to 1 (worst). We also present D-calibration results (Haider et al., 2020) in Appendix C.8.

Our evaluation employs 3-fold cross-validation with three random seeds to provide comprehensive uncertainty quantification. We report results using dual averaging perspectives. *Fold-averaged results* are computed by averaging metrics across folds within each seed, then reporting the standard deviation (std) across seeds, which captures variance from random initialization and stochastic training. *Seed-averaged results* are computed by averaging metrics across seeds within each fold and then reporting the standard deviation over folds, which captures variance from data partitioning.

4.2. Ablation Study

To isolate the contributions of GRAFT's key architectural components, we evaluate three model variants. *Full GRAFT* represents the complete model with both stochastic gates and residual MLP. The *No STG* variant removes stochastic feature selection while retaining the

residual MLP, using all input features without gating. The *Linear Only* variant removes both gates and MLP, reducing to a pure linear AFT model. We compare these variants by augmenting each dataset with additional noise covariates. For a dataset with p original features, we add kp noise features with multipliers $k \in \{3, 5, 7, 10\}$, where each noise feature is independently sampled from a standard Gaussian distribution. This experimental design allows us to quantify the contribution of the stochastic gates (by comparing *Full GRAFT* to *No STG*) and the residual MLP (by comparing *No STG* to *Linear Only*) to the overall discrimination and calibration performance.

4.3. Noise Robustness Analysis

We conduct a comprehensive noise robustness analysis by stress-testing all six models with additional noise covariates. For a dataset with p original features, we augment it with kp additional noise features at multipliers $k \in \{3, 5, 7, 10\}$, where each noise feature is independently sampled from a Student's t-distribution with 2 degrees of freedom. This distribution has infinite variance and generates extreme outliers, representing a more challenging and realistic scenario than Gaussian noise to assess feature selection robustness. This simulates the presence of spurious biomarkers or measurement artifacts in the datasets. By comparing all six models under identical noise conditions—where only the original p features are truly predictive, we can directly evaluate the effectiveness of GRAFT's stochastic gating mechanism relative to the implicit regularization strategies employed by classical models (L2 penalties in CoxPH and Weibull, bootstrap aggregation in RSF) and the vulnerabilities of deep learning approaches without explicit feature selection (DeepHit, DeepSurv). For a complete gate activation analysis and ablation study, see Appendix C.7 and C.4.

4.4. Imputation Neighborhood Sensitivity Analysis

To assess the robustness of GRAFT's local KM imputation to neighborhood construction choices, we conduct a sensitivity analysis over three design dimensions: *i*) the number of observed events required within a neighborhood $k \in \{5, 10, 20, 30\}$; *ii*) the distance metric used to identify nearest neighbors, Euclidean (Euc) or Mahalanobis (Mah); and *iii*) the neighborhood composition, where Events-Only (EO) restricts the KM estimator to neighbors with observed events only, while Events+Censored (EC) includes all neighbors regardless of censoring status. All other hyperparameters, the model architecture and the training procedure, are maintained as in Section 5.1. Each configuration is evaluated using the same 3-fold cross-validation with 3 random seeds as the main experiment, and results are reported as mean C-index and IBS across runs.

5. Results and Discussion

5.1. Baseline Comparison and Loss to C-index Alignment

Table 1 shows that GRAFT outperforms baselines in five of six datasets in C-index and IBS. In low-censoring datasets, GRAFT demonstrates clear advantages. For example, on SUPPORT, GRAFT achieves C-index 0.6143 and IBS 0.1891, outperforming all baselines including DeepSurv (0.6071) and RSF (0.6057). A similar pattern can be observed in GBSG and METABRIC. In high-censoring datasets, GRAFT maintains strong performance despite

Table 1: Comparison of GRAFT and baselines across six datasets. Values reported as Mean (Fold-Std, Seed-Std) based on 3-fold cross-validation with 3 random seeds. Fold-averaged standard deviation represents variance from model initialization; seed-averaged standard deviation represents variance from data splitting. Best values per dataset in **bold**.

Metric	Dataset	GRAFT	CoxPH	Weibull	DeepHit	DeepSurv	RSF
Low Censoring Datasets							
C-Index (higher)	GBSG	**0.6730** (0.0008, 0.0022)	0.6628 (0.0015, 0.0049)	0.6630 (0.0015, 0.0051)	0.6592 (0.0071, 0.0047)	0.6629 (0.0024, 0.0050)	0.6719 (0.0025, 0.0025)
	METABRIC	**0.6460** (0.0016, 0.0046)	0.6332 (0.0020, 0.0097)	0.6333 (0.0020, 0.0099)	0.6211 (0.0050, 0.0199)	0.6280 (0.0040, 0.0052)	0.6431 (0.0016, 0.0100)
	SUPPORT	**0.6143** (0.0010, 0.0005)	0.5686 (0.0011, 0.0041)	0.5675 (0.0011, 0.0041)	0.5737 (0.0058, 0.0023)	0.6071 (0.0016, 0.0053)	0.6057 (0.0005, 0.0048)
IBS (lower)	GBSG	**0.1760** (0.0002, 0.0012)	0.1814 (0.0003, 0.0015)	0.1821 (0.0003, 0.0017)	0.1828 (0.0006, 0.0004)	0.1818 (0.0007, 0.0023)	0.1767 (0.0006, 0.0006)
	METABRIC	0.1630 (0.0007, 0.0045)	**0.1628** (0.0006, 0.0041)	0.1636 (0.0007, 0.0037)	0.1730 (0.0007, 0.0059)	0.1728 (0.0024, 0.0042)	0.1654 (0.0004, 0.0052)
	SUPPORT	**0.1891** (0.0005, 0.0018)	0.2059 (0.0003, 0.0019)	0.2071 (0.0003, 0.0019)	0.2021 (0.0005, 0.0020)	0.1956 (0.0005, 0.0021)	0.1902 (0.0001, 0.0015)
High Censoring Datasets							
C-Index (higher)	NWTCO	**0.7173** (0.0012, 0.0055)	0.7104 (0.0005, 0.0049)	0.7113 (0.0006, 0.0047)	0.6952 (0.0086, 0.0045)	0.7101 (0.0011, 0.0075)	0.6921 (0.0025, 0.0092)
	FLCHAIN	**0.7965** (0.0004, 0.0010)	0.7689 (0.0378, 0.0400)	0.7954 (0.0004, 0.0025)	0.7695 (0.0018, 0.0181)	0.7895 (0.0003, 0.0019)	0.7917 (0.0004, 0.0047)
	AIDS	0.7301 (0.0039, 0.0244)	**0.7635** (0.0080, 0.0181)	0.7633 (0.0077, 0.0191)	0.7012 (0.0263, 0.0113)	0.6996 (0.0193, 0.0104)	0.7593 (0.0025, 0.0082)
IBS (lower)	NWTCO	**0.1016** (0.0003, 0.0007)	0.1044 (0.0000, 0.0015)	0.1064 (0.0000, 0.0017)	0.1042 (0.0011, 0.0016)	0.1073 (0.0002, 0.0016)	0.1043 (0.0003, 0.0026)
	FLCHAIN	**0.0919** (0.0001, 0.0004)	0.0961 (0.0056, 0.0060)	0.0927 (0.0003, 0.0007)	0.0972 (0.0005, 0.0016)	0.0966 (0.0013, 0.0008)	0.0928 (0.0003, 0.0009)
	AIDS	0.0572 (0.0004, 0.0032)	**0.0556** (0.0003, 0.0031)	0.0557 (0.0003, 0.0032)	0.0692 (0.0062, 0.0039)	0.0649 (0.0037, 0.0012)	0.0559 (0.0006, 0.0031)

sparse event information. In FLCHAIN (72% censoring), GRAFT achieves a C-index of 0.7965 along with the best IBS of 0.0919. The consistent superiority over classical and deep learning models validates GRAFT's hybrid architecture and training objective, providing strong empirical evidence for its utility in survival analysis.

However, GRAFT shows a clear limitation on AIDS (92% censoring), achieving C-index 0.7301 versus CoxPH's 0.7635. With only approximately 92 observed events, GRAFT's local neighborhoods contain insufficient information for stable KM curve estimation, and the added model complexity is not justified. This suggests a boundary condition: when event rates fall below approximately 10%, practitioners should consider simpler parametric models such as CoxPH. DeepHit and DeepSurv perform competitively in some datasets but lack integrated feature selection, making them vulnerable in high-dimensional, high noise settings (Section 5.3). The dual standard deviation reporting reveals GRAFT's stable convergence (fold-averaged std: 0.0002-0.0039 on the C-index) despite jointly optimizing gates, MLP parameters, and linear coefficients. Seed-averaged standard deviations (0.0005-0.0244) reflect competitive data partitioning sensitivity, with AIDS showing the highest variance due to extreme censoring.

To further validate that GRAFT's soft-Spearman loss is a reliable proxy for the target evaluation metric, we track both the training loss and the test C-index across all epochs for each of the runs of GRAFT in Table 1 (3 seeds $\times$ 3 folds) per dataset. Figure 1 shows the mean $\pm$ standard deviation of both curves on all six benchmarks. In every case, the loss decreases while the test C-index increases in tandem, confirming that minimizing the differentiable ranking objective consistently improves concordance. The tight standard deviation bands across runs further demonstrate stable convergence. The one notable exception is AIDS (92% censoring), where the loss and C-index curves exhibit higher variance. This is consistent with the quantitative results in Table 1: with only $\sim$92 observed events, the local KM neighborhoods lack sufficient event information to produce stable imputation targets, weakening the gradient signal and limiting the loss's ability to reliably guide the model toward better concordance. This reinforces our recommendation that practitioners consider simpler parametric alternatives when event rates fall below 10%.

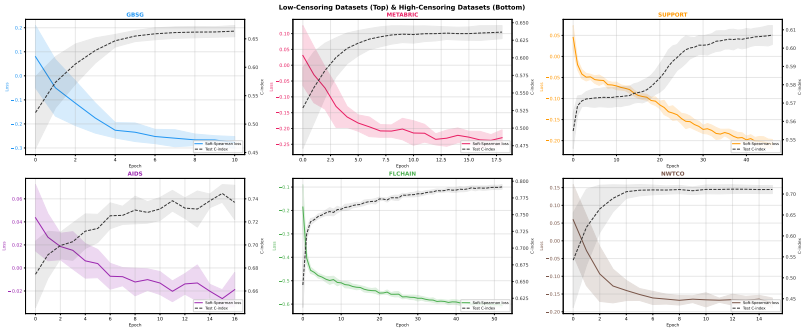

Figure 1: GRAFT training dynamics across all six datasets. Each panel shows the mean ± standard deviation of the soft-Spearman loss and test C-index over training epochs across all runs.

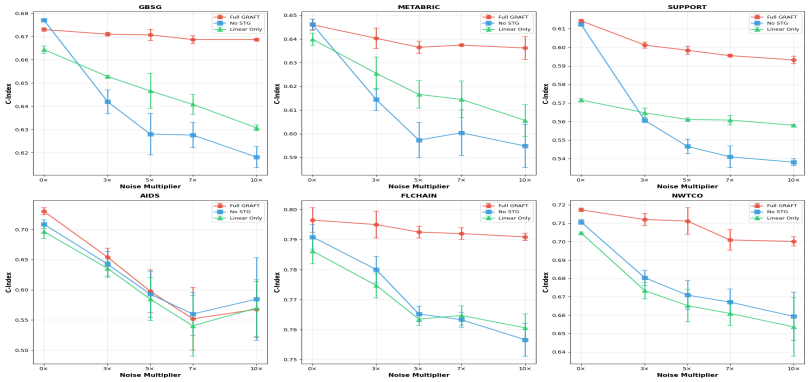

Figure 2: Ablation study showing C-index degradation under increasing Gaussian noise for three GRAFT variants. Error bars represent standard deviation across 3 random seeds.

5.2. Ablation Study

Figure 2 shows the C-index metrics across noise levels for the three GRAFT variants in all six datasets. The results reveal a clear performance hierarchy that validates our architectural design choices and quantifies the contribution of each component. At baseline (0x noise), *Full GRAFT* and *No STG* perform similarly on GBSG with a C-index of 0.673 and 0.676, respectively, while *Linear Only* achieves a lower C-index of 0.665. As noise increases, performance gaps widen dramatically. At 10x noise, *Full GRAFT* maintains a C-index of 0.669 (0.6% drop from baseline), while *No STG* drops to 0.618 (8.5% drop) and "Linear Only" drops to 0.631 (5.1% drop). Similar patterns emerge on METABRIC, SUPPORT, FLCHAIN, and NWTCO.

Interestingly, *No STG* often performs worse than *Linear Only* at high noise levels, suggesting that the flexibility of the MLP becomes harmful when operating on noisy features without protective filtering. This interaction effect underscores GRAFT's integrated design: gates and MLP work synergistically, with gates enabling the MLP to focus on truly informative features rather than fitting to noise. Dataset-specific patterns are also informative. AIDS shows the smallest performance gaps, suggesting that high-censoring limits the benefit of complex modeling. The calibration results shown in Appendix C.1

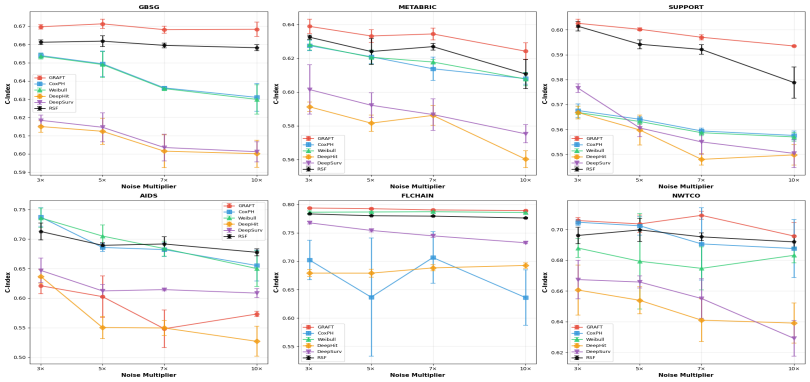

Figure 3: Comparison of C-index for all six models under heavy-tailed Student's t noise ($df = 2$). Error bars represent standard deviation across 3 random seeds.

indicate that IBS follows a similar pattern, demonstrating that feature selection improves both discrimination and calibration.

5.3. Noise Robustness Analysis

Figure 3 presents the C-index metrics for all six models under heavy-tailed Student's t noise ($df = 2$) across increasing noise multipliers. This experiment reveals fundamental differences in how each model handles spurious high-dimensional features and provides direct evidence of GRAFT's superior robustness. On GBSG, GRAFT's C-index decreases minimally from 0.670 at 3x noise to 0.669 at 10x noise (0.15% degradation), while DeepSurv drops from 0.619 to 0.601 (2.9% degradation). On METABRIC, SUPPORT, FLCHAIN and NWTCO, GRAFT remains relatively stable across all noise levels. RSF emerges as the second best model in terms of robustness across different noise levels, with bootstrap sampling likely providing some implicit noise resistance. However, RSF's performance still degrades more than GRAFT. On SUPPORT, RSF's C-index drops from 0.602 at 3x noise to 0.579 at 10x noise (3.8% degradation), while GRAFT only drops from 0.604 to 0.594 (1.6% degradation). The contrast between approaches is instructive. Classical models (CoxPH, Weibull) show moderate robustness but lack explicit feature selection, leading to gradual degradation as noise accumulates. RSF shows better robustness, though inferior to GRAFT's explicit feature gating. Deep learning models exhibit severe vulnerabilities: DeepSurv and DeepHit show catastrophic degradation, which is likely due to overfitting to noise patterns without mechanisms to distinguish signal from spurious features.

The key insight is that GRAFT's stochastic gates provide explicit, learnable noise filtering superior to implicit regularization. GRAFT learns a sparse binary mask that selectively suppresses uninformative features through differentiable L0 regularization. The nearly flat performance curves demonstrate that GRAFT successfully identifies true signal features and ignores injected noise, even when noise features outnumber signal features 10 to 1. Dataset-specific patterns reveal informative variations: FLCHAIN shows remarkable stability across multiple models, suggesting strong signal quality despite high censoring. Moreover, GBSG, METABRIC, and SUPPORT show clear model separation with GRAFT's advantages most pronounced; while AIDS reinforces that extreme censoring (92%) favors simpler parametric

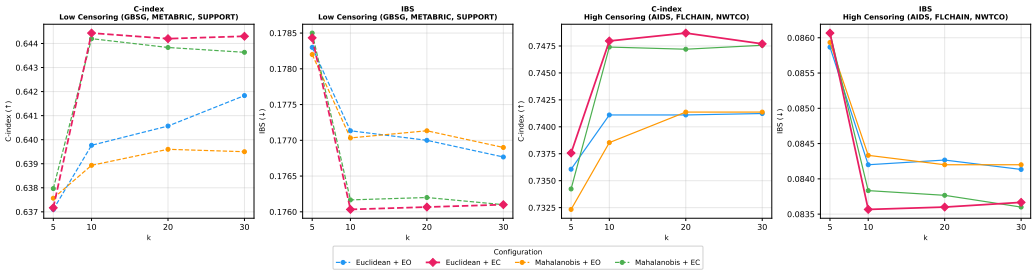

Figure 4: Sensitivity of C-index (left) and IBS (right) to neighborhood construction choices across all six datasets. Each point is the mean across all runs.

assumptions. The calibration results shown in Appendix C.1 confirm that IBS follows similar trends, with GRAFT maintaining stable calibration performance, while competing models show increasing calibration errors under noise.

5.4. Imputation Neighborhood Sensitivity Analysis

Figure 4 reports the C-index and IBS for all neighborhood construction configurations. Two findings emerge consistently in all six datasets. First, performance stabilizes rapidly once k reaches 10, with negligible changes at 20 and 30. This validates our choice of k in the main experiments: requiring fewer events (5) yields unstable KM estimates with marginally worse discrimination, while requiring more (20, 30) offers no measurable benefit. Second, the distance metric and the neighborhood composition have minimal impact on performance. Both Euclidean and Mahalanobis distances produce near-identical results across all configurations with Euclidean slightly outperforming Mahalanobis. EC consistently outperforms EO, indicating that including censored neighbors in the KM fit is essential to avoid biased estimates. These results demonstrate that GRAFT's imputation mechanism is robust to neighborhood construction choices. Detailed results are presented in Appendix C.2.

6. Conclusions

We introduced GRAFT, a novel survival analysis model built around three core contributions: the explicit decoupling of prognostic ranking from survival calibration, a hybrid architecture that combines linear and non-linear components trained via stochastic imputation and a C-index-aligned ranking loss, and integrated feature selection via stochastic gates. Our experiments demonstrate that GRAFT outperforms baselines in both ranking and calibration. The ablation study further confirms that the architectural components contribute synergistically, with stochastic gates providing robust performance preservation under extreme noise conditions.

Limitations GRAFT is particularly well-suited for high-dimensional biomedical applications where automated feature selection is critical. However, the model shows limitations under extreme censoring ($\geq 92\%$), where simpler parametric models may be preferable due to insufficient event information for reliable local neighborhood estimation. Future work could explore adaptive architectures for extreme censoring regimes and extensions to competing risk scenarios.

References

Ahtisham Fazeel Abbasi, Muhammad Nabeel Asim, Sheraz Ahmed, Sebastian Vollmer, and Andreas Dengel. Survival prediction landscape: an in-depth systematic literature review on activities, methods, tools, diseases, and databases. *Frontiers in Artificial Intelligence*, 7:1428501, 2024.

Rudolf Beran. Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley, 1981.

Michael J Best, Nilotpal Chakravarti, and Vasant A Ubhaya. Minimizing separable convex functions subject to simple chain constraints. *SIAM Journal on Optimization*, 10(3): 658–672, 2000.

Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020.

Norman E Breslow and Nilanjan Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):457–468, 1999.

Lukas Burk, John Zobolas, Bernd Bischl, Andreas Bender, Marvin N Wright, and Raphael Sonabend. A large-scale neutral comparison study of survival models on low-dimensional data. *arXiv preprint arXiv:2406.04098*, 2024.

TG Clark, MJ Bradburn, SB Love, and DG Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

Dorota M Dabrowska. Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics*, pages 181–197, 1987.

Dorota M Dabrowska. Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Annals of Statistics*, 17(3):1157–1167, 1989.

Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

Lore Dirick, Gerda Claeskens, and Bart Baesens. Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, 68(6): 652–665, 2017.

Angela Dispenzieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton III, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, pages 517–523. Elsevier, 2012.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85): 1–63, 2020. URL http://jmlr.org/papers/v21/18-772.html.

Scott M Hammer, Kathleen E Squires, Michael D Hughes, Janet M Grimes, Lisa M Demeter, Judith S Currier, Joseph J Eron Jr, Judith E Feinberg, Henry H Balfour Jr, Lawrence R Deyton, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, 337(11):725–733, 1997.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

Weiyi He. Review for dynamic prediction in clinical survival analysis. *arXiv preprint arXiv:2311.15743*, 2023.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2002.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Aziz Kassani, Mohsen Niazi, Jafar Hassanzadeh, and Rostam Menati. Survival analysis of drug abuse relapse in addiction treatment centers. *International journal of high risk behaviors & addiction*, 4(3):e23402, 2015.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1): 1–12, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

DY Lin. On the breslow estimator. *Lifetime data analysis*, 13(4):471–480, 2007.

Joanne Lynn, Frank E. Harrell, Felicia Cohn, Douglas Wagner, and Alfred F. Connors. Prognoses of seriously ill hospitalized patients on the days before death: implications for patient care and public policy. *New Horizons*, 5(1):56–61, 1997.

William Q Meeker, Luis A Escobar, and Francis G Pascual. *Statistical methods for reliability data*. John Wiley & Sons, 2014.

Andrew Miller, Nick Foti, Alexander D'Amour, and Ryan P Adams. Reducing reparameterization gradient variance. *Advances in Neural Information Processing Systems*, 30, 2017.

Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

Rinku Saikia and Manash Pratim Barman. A review on accelerated failure time models. *International Journal of Statistics and Systems*, 12(2):311–322, 2017.

Martin Schumacher, G Bastert, H Bojar, K Hübner, M Olschewski, W Sauerbrei, C Schmoor, C Beyerle, RL Neumann, and HF Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Ryan J Tibshirani, Holger Hoefling, and Robert Tibshirani. Nearly-isotonic regression. *Technometrics*, 53(1):54–61, 2011.

Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. *Proceedings of the 37th International Conference on Machine Learning*, pages 10648–10659, 2020.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(11), 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix A. Broader Impact

This paper presents work whose goal is to advance the field of machine learning in healthcare, specifically survival analysis for time-to-event prediction. Our work has direct applications in healthcare, where survival models inform critical clinical decisions including treatment planning, risk stratification, and resource allocation. While improved prediction accuracy can lead to better patient outcomes, we acknowledge several important considerations.

First, survival models trained on historical data may perpetuate existing biases in healthcare delivery if demographic disparities are present in training data. Practitioners should carefully audit model predictions across patient subgroups before clinical deployment. Second, the interpretability provided by our linear AFT component and feature selection mechanism, while beneficial for model understanding, should not substitute for clinical judgment—models should augment rather than replace physician decision-making. Third, the automatic feature selection capability, though valuable for identifying relevant biomarkers, requires validation that selected features reflect true biological mechanisms rather than spurious correlations.

We emphasize that GRAFT, like all machine learning models for healthcare, requires rigorous clinical validation, fairness auditing, and regulatory approval before deployment in medical settings. The model's performance limitations under extreme censoring ($\geq 92\%$) highlight the importance of understanding applicability boundaries. Responsible use requires ongoing monitoring, transparent communication of uncertainty, and integration within broader clinical workflows that maintain human oversight.

Appendix B. Hyperparameters

For reproducibility purposes, all hyperparameters used for GRAFT and baseline models in the experiments are specified below in Table 2.

Table 2: Hyperparameters for all models used in experiments. All deep learning models use comparable hyperparameters to ensure fair comparison.

Model	Hyperparameter	Value
GRAFT	Hidden dimension (d_h)	32
	Learning rate	10^{-3}
	Weight decay (α_{L2})	10^{-4}
	L0 regularization (λ_{L0})	0.01
	Dropout rate	0.2
	STG noise std (σ)	0.5
	Soft-rank smoothing (τ)	0.1
	Batch size (m)	64
	Monte Carlo samples (M)	5
	Training epochs (E)	1000 with early stopping
	Patience	10
	Min obs events per neighborhood (k)	10
	Distance metric	Euclidean
	Optimizer	Adam
CoxPH	L2 penalty (penalizer)	0.01
Weibull AFT	L2 penalty (penalizer)	0.01
DeepHit	Number of duration bins	100
	Hidden layers	[32, 32]
	Batch normalization	True
	Dropout rate	0.2
	Learning rate	0.01
	Optimizer	Adam
	Alpha (ranking loss weight)	0.3
	Sigma (smoothing parameter)	0.1
	Batch size	64
	Training epochs	1000 with early stopping
	Patience	10
DeepSurv	Hidden layers	[32, 32]
	Batch normalization	True
	Dropout rate	0.2
	Learning rate	0.01
	Optimizer	Adam
	Output bias	False
	Batch size	64
	Training epochs	1000 with early stopping
	Patience	10
RSF	Number of trees	100
	Min samples split	10
	Min samples leaf	15
	Max features	sqrt

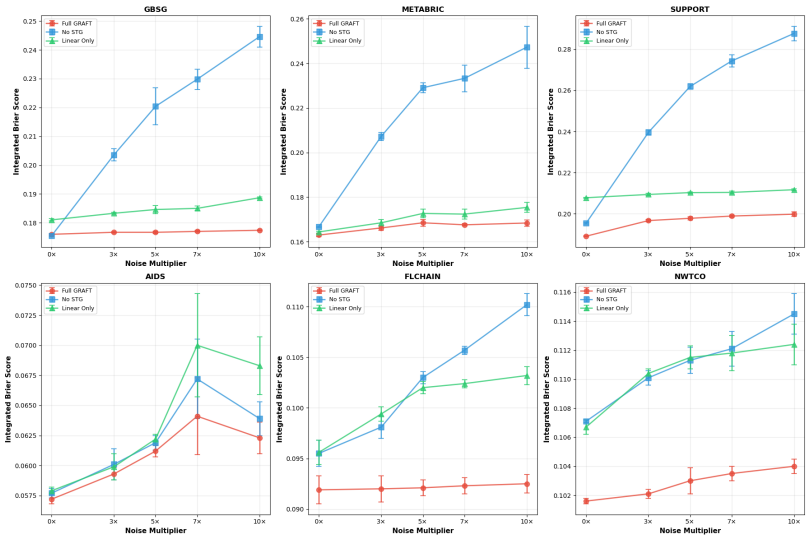

Figure 5: Ablation study showing IBS degradation under increasing Gaussian noise for three GRAFT variants across six datasets. "Full GRAFT" (red) maintains stable performance across noise levels. Error bars represent standard deviation across 3 random seeds.

Appendix C. Additional Results and Experiments

C.1. IBS Results for Ablation and Noise Experiment

The ablation study (Sec 5.2) and noise robustness analysis (Sec 5.3) presented C-index results. Figures 5 and 6 present the corresponding IBS results across all six datasets. The calibration metrics follow the same qualitative pattern. In Figure 5 *Full GRAFT* maintains stable IBS as noise increases, while the *No STG* variant shows degradation due to overfitting on irrelevant features, and the *Linear Only* variant exhibits limited flexibility. Figure 6 demonstrates GRAFT's superior discrimination performance under increasing noise. These results confirm that GRAFT's stochastic gates improve both discrimination and calibration in high-dimensional, noisy settings.

C.2. Full Results for Imputation Neighbourhood Sensitivity Analysis

Tables 3 and 4 report the complete C-index and IBS results for all neighbourhood construction configurations across all six datasets. Each cell reports the mean and standard deviation across all runs (3 seeds × 3 folds). The configurations vary along three dimensions: the minimum number of observed events required within a neighbourhood (`min_events` $\in \{5, 10, 20, 30\}$), the distance metric (Euclidean or Mahalanobis), and the neighbourhood composition, where EO (Events-Only) restricts the local Kaplan-Meier estimator to neighbours with observed events, while EC (Events+Censored) includes all neighbours regardless of censoring status. All other hyperparameters are fixed at the values used in the main experiments. The configuration used in the paper (`k=10`, Euclidean, EC) is marked with †.

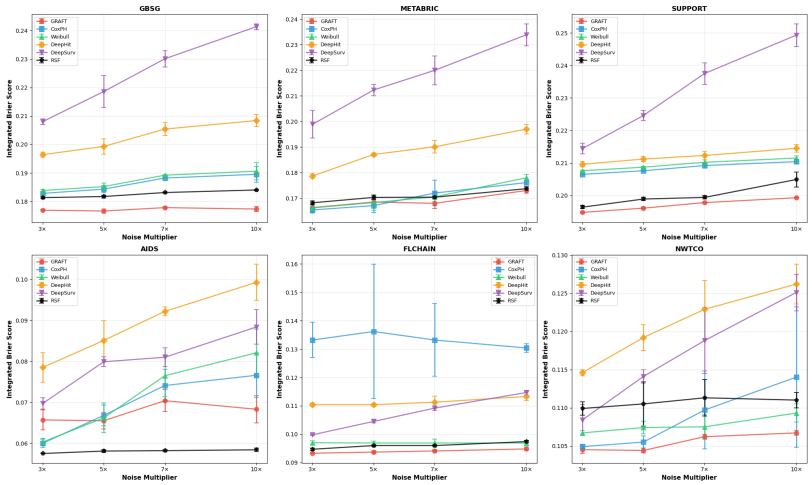

Figure 6: Noise robustness comparison showing IBS under increasing Student's t noise ($df = 2$) for all six models across six datasets. GRAFT (red) maintains the flattest performance curves, demonstrating superior robustness to heavy-tailed noise. Error bars represent standard deviation across 3 random seeds.

Table 3: Harrell's C-index (mean ± std across all runs, higher is better) for all neighbourhood construction configurations across six datasets. Low-censoring: GBSG (43%), METABRIC (42%), SUPPORT (32%). High-censoring: AIDS (92%), FLCHAIN (72%), NWTCO (86%). Best value per dataset in **bold**. † Paper configuration. EO = Events-Only; EC = Events+Censored.

k	Dist.	Comp.	GBSG	METABRIC	SUPPORT	AIDS	FLCHAIN	NWTCO
Low censoring datasets: GBSG, METABRIC, SUPPORT — High censoring datasets: AIDS, FLCHAIN, NWTCO								
5	Euc	EO	0.6649 (0.0021)	0.6366 (0.0037)	0.6098 (0.0011)	0.7105 (0.0052)	0.7859 (0.0025)	0.7118 (0.0014)
5	Euc	EC	0.6662 (0.0011)	0.6371 (0.0038)	0.6082 (0.0019)	0.7111 (0.0069)	0.7889 (0.0032)	0.7127 (0.0017)
5	Mah	EO	0.6662 (0.0024)	0.6368 (0.0028)	0.6097 (0.0013)	0.6960 (0.0071)	0.7887 (0.0024)	0.7123 (0.0018)
5	Mah	EC	0.6669 (0.0019)	0.6378 (0.0042)	0.6092 (0.0016)	0.7037 (0.0064)	0.7860 (0.0027)	0.7130 (0.0016)
10	Euc	EO	0.6693 (0.0015)	0.6390 (0.0015)	0.6110 (0.0013)	0.7173 (0.0045)	0.7897 (0.0010)	0.7163 (0.0017)
10	Euc	EC†	**0.6730** (0.0008)	0.6460 (0.0016)	**0.6143** (0.0010)	0.7301 (0.0039)	0.7965 (0.0004)	0.7173 (0.0012)
10	Mah	EO	0.6690 (0.0008)	0.6385 (0.0014)	0.6093 (0.0012)	0.7113 (0.0060)	0.7884 (0.0009)	0.7159 (0.0013)
10	Mah	EC	0.6728 (0.0010)	**0.6461** (0.0017)	0.6137 (0.0017)	0.7292 (0.0047)	0.7955 (0.0005)	0.7175 (0.0015)
20	Euc	EO	0.6694 (0.0012)	0.6398 (0.0017)	0.6125 (0.0061)	0.7176 (0.0061)	0.7898 (0.0014)	0.7159 (0.0013)
20	Euc	EC	0.6728 (0.0010)	0.6458 (0.0014)	0.6140 (0.0009)	**0.7328** (0.0036)	**0.7966** (0.0007)	0.7167 (0.0012)
20	Mah	EO	0.6692 (0.0011)	0.6392 (0.0014)	0.6104 (0.0046)	0.7182 (0.0057)	0.7899 (0.0013)	0.7160 (0.0013)
20	Mah	EC	0.6718 (0.0009)	0.6459 (0.0010)	0.6138 (0.0018)	0.7280 (0.0041)	0.7963 (0.0004)	0.7173 (0.0012)
30	Euc	EO	0.6692 (0.0007)	0.6428 (0.0017)	0.6135 (0.0012)	0.7181 (0.0032)	0.7897 (0.0008)	0.7159 (0.0011)
30	Euc	EC	**0.6730** (0.0012)	0.6458 (0.0016)	0.6141 (0.0011)	0.7295 (0.0047)	0.7964 (0.0006)	0.7172 (0.0015)
30	Mah	EO	0.6693 (0.0009)	0.6394 (0.0010)	0.6098 (0.0015)	0.7189 (0.0035)	0.7892 (0.0014)	0.7160 (0.0010)
30	Mah	EC	0.6728 (0.0008)	0.6451 (0.0012)	0.6130 (0.0019)	0.7289 (0.0053)	0.7962 (0.0004)	**0.7176** (0.0014)

C.3. Alternative Method for Survival Curve Estimation: Isotonic Regression

In Section 3.6, we employ Cox Proportional Hazards (CoxPH) to calibrate GRAFT's prognostic scores into survival curves. While this approach is computationally efficient and provides smooth, well-calibrated survival estimates, it introduces a potential methodological tension: Cox calibration step explicitly assumes that hazard ratios remain constant over time. To address this, we evaluate an alternative non-parametric calibration method based

Table 4: Integrated Brier Score (mean ± std across all runs, lower is better) for all neighbourhood construction configurations across six datasets. Calibration obtained via post-hoc 1D CoxPH on GRAFT scores, identical to the main experiments. Best value per dataset in **bold**. † Paper configuration. EO = Events-Only; EC = Events+Censored.

k	Dist.	Comp.	GBSG	METABRIC	SUPPORT	AIDS	FLCHAIN	NWTCO
Low censoring datasets: GBSG, METABRIC, SUPPORT — High censoring datasets: AIDS, FLCHAIN, NWTCO								
5	Euc	EO	0.1783 (0.0008)	0.1632 (0.0013)	0.1934 (0.0024)	0.0591 (0.0017)	0.0951 (0.0019)	0.1034 (0.0014)
5	Euc	EC	0.1784 (0.0006)	0.1634 (0.0017)	0.1935 (0.0022)	0.0591 (0.0016)	0.0954 (0.0020)	0.1037 (0.0011)
5	Mah	EO	0.1780 (0.0006)	0.1632 (0.0020)	0.1934 (0.0025)	0.0592 (0.0015)	0.0954 (0.0026)	0.1032 (0.0013)
5	Mah	EC	0.1785 (0.0006)	0.1633 (0.0014)	0.1937 (0.0022)	0.0592 (0.0017)	0.0951 (0.0026)	0.1039 (0.0012)
10	Euc	EO	0.1764 (0.0006)	0.1631 (0.0004)	0.1919 (0.0008)	0.0581 (0.0006)	0.0922 (0.0002)	0.1023 (0.0005)
10	Euc	EC†	**0.1760** (0.0002)	0.1630 (0.0007)	**0.1891** (0.0005)	0.0572 (0.0004)	**0.0919** (0.0001)	**0.1016** (0.0003)
10	Mah	EO	0.1765 (0.0006)	0.1631 (0.0006)	0.1915 (0.0012)	0.0583 (0.0007)	0.0925 (0.0006)	0.1022 (0.0004)
10	Mah	EC	**0.1760** (0.0005)	0.1628 (0.0007)	0.1897 (0.0007)	0.0575 (0.0006)	0.0921 (0.0007)	0.1019 (0.0005)
20	Euc	EO	0.1761 (0.0009)	0.1630 (0.0007)	0.1919 (0.0005)	0.0582 (0.0009)	0.0923 (0.0008)	0.1023 (0.0006)
20	Euc	EC	0.1761 (0.0009)	0.1629 (0.0006)	0.1892 (0.0004)	0.0572 (0.0008)	**0.0919** (0.0004)	0.1017 (0.0003)
20	Mah	EO	0.1762 (0.0007)	0.1631 (0.0012)	0.1921 (0.0009)	0.0583 (0.0007)	0.0923 (0.0005)	0.1020 (0.0005)
20	Mah	EC	0.1762 (0.0004)	**0.1626** (0.0010)	0.1898 (0.0006)	0.0574 (0.0011)	0.0920 (0.0004)	0.1019 (0.0004)
30	Euc	EO	0.1763 (0.0004)	0.1629 (0.0006)	0.1911 (0.0006)	0.0581 (0.0007)	0.0921 (0.0007)	0.1022 (0.0006)
30	Euc	EC	0.1761 (0.0005)	0.1629 (0.0008)	0.1893 (0.0009)	**0.0571** (0.0008)	0.0920 (0.0005)	0.1019 (0.0007)
30	Mah	EO	0.1762 (0.0004)	0.1630 (0.0005)	0.1915 (0.0010)	0.0583 (0.0006)	0.0922 (0.0006)	0.1021 (0.0005)
30	Mah	EC	0.1763 (0.0003)	0.1628 (0.0011)	0.1892 (0.0007)	0.0573 (0.0007)	**0.0919** (0.0003)	**0.1016** (0.0004)

on isotonic regression. For a grid of $K = 50$ discrete time points spanning the test set's time range, we fit one isotonic regression at each time point. At a given time point t_k (for $k = 1, \ldots, K$), we include subjects with known survival status at t_k: those with $t_i > t_k$ (who survived past t_k), and those with $\delta_i = 1$ and $t_i \leq t_k$ (whose event occurred by t_k). Subjects with $\delta_i = 0$ and $t_i \leq t_k$ are excluded, as their survival status at t_k is ambiguous. We construct binary survival indicators as targets: $y_i^{(k)} = 1$ for subjects with $t_i > t_k$ and $y_i^{(k)} = 0$ for subjects with $\delta_i = 1$ and $t_i \leq t_k$. Isotonic regression then learns a function from GRAFT scores to these binary targets, yielding calibrated survival probability estimates $\hat{S}(t_k|s_i) \in [0, 1]$ where s_i denotes the GRAFT score for subject i. Survival curves are constructed by linear interpolation between the $K = 50$ calibrated probability estimates.

This isotonic approach offers two potential advantages over Cox: (1) it makes no parametric assumptions about the baseline hazard or proportional hazards, and (2) each time point learns an independent score-to-probability mapping, allowing the relationship to vary over time (capturing non-proportional hazards effects if present). However, it also has potential drawbacks: (1) reduced sample size at each time point due to censoring, especially at later times, and (2) interpolation artifacts between discrete time points. Table 5 compares Integrated Brier Scores (IBS) for GRAFT calibrated with Cox vs isotonic regression across all six benchmark datasets. As shown in Table 5, both calibration methods achieve competitive performance. Cox calibration demonstrates better mean IBS on five of six datasets and lower standard deviations across both fold-averaged and seed-averaged metrics. These properties—competitive calibration accuracy, lower variance, and computational efficiency motivated our choice of Cox calibration for GRAFT. However, the comparable performance of isotonic regression demonstrates that practitioners may opt for isotonic regression when non-proportional hazards effects are strongly suspected or when fully non-parametric calibration is desired.

Table 5: Comparison of calibration methods: Cox PH *vs.* Isotonic Regression. Lower IBS is better. Values reported as Mean (Fold-Std, Seed-Std).

Dataset	GRAFT	GRAFT w/ Isotonic
GBSG	**0.1760** (0.0002, 0.0012)	0.1782 (0.0004, 0.0014)
METABRIC	0.1630 (0.0007, 0.0045)	**0.1627** (0.0009, 0.0048)
SUPPORT	**0.1891** (0.0005, 0.0018)	0.1902 (0.0006, 0.0024)
NWTCO	**0.1016** (0.0003, 0.0007)	0.1032 (0.0006, 0.0010)
FLCHAIN	**0.0919** (0.0001, 0.0004)	0.0927 (0.0003, 0.0007)
AIDS	**0.0572** (0.0004, 0.0032)	0.0602 (0.0009, 0.0045)

Table 6: Ablation study comparing gating mechanisms across six datasets under increasing noise conditions. Values reported as Mean (Fold-Std, Seed-Std) based on 3-fold cross-validation with 3 random seeds. Noise levels: 0x, 5x, 10x. Best values per dataset and noise level in bold.

Metric	Dataset	Noise	STG	Sigmoid	REINFORCE
Low Censoring Datasets					
C-index (higher)	GBSG	0x	**0.6730** (0.0008, 0.0022)	0.6691 (0.0020, 0.0035)	0.6722 (0.0015, 0.0028)
		5x	**0.6713** (0.0026, 0.0026)	0.6583 (0.0040, 0.0055)	0.6680 (0.0030, 0.0042)
		10x	0.6683 (0.0019, 0.0039)	0.6450 (0.0050, 0.0080)	**0.6690** (0.0038, 0.0065)
	METABRIC	0x	**0.6460** (0.0016, 0.0046)	0.6421 (0.0030, 0.0065)	0.6350 (0.0025, 0.0055)
		5x	0.6332 (0.0038, 0.0038)	**0.6334** (0.0060, 0.0075)	0.6280 (0.0052, 0.0065)
		10x	**0.6242** (0.0035, 0.0050)	0.5990 (0.0075, 0.0095)	0.6184 (0.0065, 0.0072)
	SUPPORT	0x	**0.6143** (0.0010, 0.0005)	0.5992 (0.0025, 0.0020)	0.6050 (0.0018, 0.0012)
		5x	**0.6002** (0.0006, 0.0006)	0.5786 (0.0020, 0.0030)	0.5959 (0.0015, 0.0028)
		10x	**0.5935** (0.0010, 0.0002)	0.5681 (0.0035, 0.0040)	0.5882 (0.0025, 0.0028)
IBS (lower)	GBSG	0x	**0.1760** (0.0002, 0.0012)	0.1761 (0.0010, 0.0022)	**0.1760** (0.0008, 0.0018)
		5x	0.1766 (0.0007, 0.0007)	0.1920 (0.0018, 0.0030)	**0.1755** (0.0012, 0.0020)
		10x	**0.1773** (0.0003, 0.0009)	0.1972 (0.0025, 0.0045)	0.1850 (0.0018, 0.0035)
	METABRIC	0x	**0.1630** (0.0007, 0.0045)	0.1682 (0.0020, 0.0065)	0.1655 (0.0015, 0.0058)
		5x	**0.1685** (0.0013, 0.0013)	0.1859 (0.0035, 0.0075)	0.1750 (0.0028, 0.0065)
		10x	**0.1730** (0.0020, 0.0011)	0.2002 (0.0045, 0.0090)	0.1824 (0.0038, 0.0075)
	SUPPORT	0x	**0.1891** (0.0005, 0.0018)	0.1901 (0.0015, 0.0035)	0.1899 (0.0012, 0.0028)
		5x	**0.1961** (0.0003, 0.0003)	0.1988 (0.0020, 0.0040)	0.1974 (0.0015, 0.0032)
		10x	**0.1993** (0.0001, 0.0001)	0.2282 (0.0030, 0.0050)	0.2080 (0.0020, 0.0040)
High Censoring Datasets					
C-index (higher)	NWTCO	0x	**0.7173** (0.0012, 0.0055)	0.7122 (0.0035, 0.0075)	0.7080 (0.0025, 0.0065)
		5x	0.7036 (0.0057, 0.0057)	0.6780 (0.0080, 0.0115)	**0.7051** (0.0070, 0.0095)
		10x	**0.6957** (0.0026, 0.0089)	0.6620 (0.0065, 0.0120)	0.6921 (0.0050, 0.0100)
	FLCHAIN	0x	**0.7965** (0.0004, 0.0010)	0.7950 (0.0018, 0.0030)	0.7947 (0.0010, 0.0022)
		5x	**0.7926** (0.0009, 0.0009)	0.7683 (0.0025, 0.0040)	0.7880 (0.0018, 0.0030)
		10x	**0.7893** (0.0002, 0.0004)	0.7591 (0.0035, 0.0055)	0.7820 (0.0025, 0.0040)
	AIDS	0x	0.7301 (0.0039, 0.0244)	0.7296 (0.0180, 0.0300)	**0.7314** (0.0155, 0.0265)
		5x	**0.6026** (0.0351, 0.0285)	0.6020 (0.0450, 0.0420)	0.6019 (0.0400, 0.0390)
		10x	0.5732 (0.0318, 0.0047)	0.5380 (0.0480, 0.0200)	**0.5823** (0.0390, 0.0180)
IBS (lower)	NWTCO	0x	**0.1016** (0.0003, 0.0007)	**0.1016** (0.0015, 0.0020)	0.1095 (0.0012, 0.0015)
		5x	**0.1044** (0.0003, 0.0003)	0.1180 (0.0015, 0.0020)	0.1095 (0.0010, 0.0015)
		10x	**0.1067** (0.0003, 0.0003)	0.1282 (0.0020, 0.0030)	0.1140 (0.0015, 0.0025)
	FLCHAIN	0x	**0.0919** (0.0001, 0.0004)	0.0938 (0.0010, 0.0015)	0.0975 (0.0008, 0.0012)
		5x	**0.0937** (0.0004, 0.0004)	0.1001 (0.0015, 0.0025)	0.0968 (0.0010, 0.0018)
		10x	**0.0948** (0.0004, 0.0004)	0.1120 (0.0025, 0.0040)	0.1015 (0.0020, 0.0030)
	AIDS	0x	**0.0572** (0.0004, 0.0032)	0.0589 (0.0040, 0.0050)	0.0579 (0.0035, 0.0045)
		5x	**0.0655** (0.0020, 0.0020)	0.0682 (0.0050, 0.0060)	0.0691 (0.0040, 0.0050)
		10x	0.0683 (0.0027, 0.0033)	0.0790 (0.0070, 0.0190)	**0.0675** (0.0050, 0.0160)

C.4. Gating Mechanism Ablation Study

To validate our choice of Stochastic Gates (STG) as the feature selection mechanism in GRAFT, we conducted an ablation study comparing three gating approaches across all six benchmark datasets across. For a dataset with p original features, we augment it with kp additional noise features at multipliers $k \in \{5, 10\}$, where each noise feature is independently sampled from a heavy-tailed Student's t-distribution with 2 degrees of freedom. The three variants tested were: *i*) **STG**: a differentiable feature selection mechanism based on continuous relaxation of Bernoulli variables Yamada et al. (2020); *ii*) **Sigmoid**: a simpler deterministic gating mechanism using learnable parameters $\alpha \in \mathbb{R}^p$, where gates are computed as $g_j = \text{sigmoid}(\alpha_j)$ for each j-th feature, without stochasticity; and *iii*) **REINFORCE**: a policy gradient approach (Williams, 1992) where binary gates are sampled from a Bernoulli distribution and updated using the REINFORCE algorithm with variance reduction and entropy regularization. All three variants use the same GRAFT architecture (linear AFT + residual MLP), the same local KM imputation procedure, and the same soft-ranking training objective—only the gating mechanism differs. Table 6 presents the results across all datasets. At 0x noise, all three gating mechanisms perform similarly; for example, in FLCHAIN, STG achieves a C-index of 0.7965, while Sigmoid and REINFORCE achieve 0.7950 and 0.7947, respectively. However, as noisy features increase, STG consistently demonstrates superior performance preservation, maintaining better discrimination (C-index) and calibration (IBS) compared to the alternatives. REINFORCE ranks second, showing moderate robustness to noise, while Sigmoid exhibits significant performance degradation under noisy conditions.

C.5. Sensitivity to Soft-Rank Smoothing Parameter

The soft-ranking operator in Section 3.5 (11) relies on a smoothing parameter $\tau > 0$ that controls the sharpness of the continuous relaxation of the rank function: small τ values yield a closer approximation to the true (non-differentiable) rank operator at the cost of sharper gradients, while large τ values produce smoother, more conservative gradients. To assess GRAFT's sensitivity to this hyperparameter, we conduct a sweep over $\tau \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$ on SUPPORT, keeping all other hyperparameters fixed at the values in Appendix B. Table 7 reports fold-averaged C-index and IBS (mean ± std across 3 seeds) for each value of τ.

Table 7: Sensitivity of GRAFT to the soft-rank smoothing parameter τ on SUPPORT. Values reported as fold-averaged mean ± std across 3 seeds. Our default $\tau = 0.1$ is marked with †.

τ	**C-index (↑)**	**IBS (↓)**
0.01	0.6086 ± 0.0058	0.1919 ± 0.0017
0.05	0.6139 ± 0.0010	0.1892 ± 0.0003
0.1 †	0.6143 ± 0.0010	0.1891 ± 0.0005
0.2	0.6143 ± 0.0010	0.1891 ± 0.0005
0.5	0.6143 ± 0.0010	0.1891 ± 0.0005
1.0	0.6143 ± 0.0010	0.1891 ± 0.0005
2.0	0.6143 ± 0.0010	0.1891 ± 0.0005

Table 8: Results on TCGA-BRCA (1,086 patients, 18,459 genes, 85.8% censoring), a $p \gg n$ genomic benchmark. Values reported as mean ± std across 3 seeds × 3 folds. CoxPH and Weibull AFT failed to converge due to Hessian inversion infeasibility at $p \gg n$. Best value per metric in **bold**.

Metric	GRAFT	CoxPH	Weibull AFT	DeepHit	DeepSurv	RSF
C-index (↑)	**0.7243** ± 0.0590	–	–	0.5718 ± 0.0453	0.5924 ± 0.0302	0.6093 ± 0.0447
IBS (↓)	**0.1826** ± 0.0259	–	–	0.1993 ± 0.0350	0.2392 ± 0.0262	0.2003 ± 0.0324

Performance is stable across $\tau \in [0.05, 2.0]$, with noticeable degradation only at $\tau = 0.01$, where overly sharp gradients induce training instability, reflected in both a lower mean C-index and higher variance across seeds. Notably, large τ values do not degrade performance on SUPPORT, which we attribute to the dataset's well-separated event times; we expect that degradation at the upper end of the range would emerge on datasets with a higher prevalence of tied event times, where excessive smoothing would blur meaningful rank distinctions. Our default value of $\tau = 0.1$ falls within the range recommended by Blondel et al. (2020) and achieves the best (or tied-best) performance.

C.6. Validation on High-Dimensional Genomic Data: TCGA-BRCA

To further evaluate GRAFT's applicability in realistic high-dimensional clinical settings, we conduct an experiment on TCGA-BRCA (The Cancer Genome Atlas Network, 2012), a breast cancer transcriptomics dataset comprising 1,086 patients, 18,459 protein-coding genes, and 85.8% censoring. This dataset constitutes a genuine $p \gg n$ regime, with approximately 724 training samples per fold against 18,459 candidate features.

Under this extreme dimensionality, both CoxPH and Weibull AFT failed to converge. With approximately 724 training samples per fold and only ∼100–150 observed events (85.8% censoring), the Hessian of the (partial) log-likelihood over $p = 18{,}459$ parameters has rank far below its dimension, making it singular (or numerically near-singular) and causing its inversion to fail during optimization. we report these as failures ("–") in Table 8. In contrast, GRAFT's stochastic gates automatically performed feature selection in this setting, opening only 623 ± 36 of the 18,459 available genes on average across runs while still achieving the strongest discrimination and calibration among all evaluated models. Table 8 reports results from 3-seed, 3-fold cross-validation. These findings support GRAFT's utility for real-world, high-dimensional biomedical applications such as genomic survival analysis.

C.7. Gate Activation Analysis Across All Datasets

To directly verify that GRAFT's stochastic gates suppress injected noise features, we analyze gate activation patterns across all six benchmark datasets under the heavy-tailed Student's t noise ($df = 2$) setting described in Section 5.3. In this experiment, a gate is considered *open* if its deterministic value in (9) exceeds a threshold of 0.5 at test time. Table 9 reports the mean number of open gates among real and injected noise features, averaged across 3 seeds × 3 folds, for noise multipliers $k \in \{0, 3, 5, 7, 10\}$.

Across low-censoring datasets (GBSG, METABRIC, SUPPORT), GRAFT achieves near or fully-complete suppression of injected noise features at every noise level, with GBSG and

Table 9: Gate activation statistics across all six datasets (threshold = 0.5), mean ± std across 3 seeds × 3 folds. *Open:Real* counts original features with open gates (out of p total); *Open:Noise* counts injected noise features with open gates (out of p_{noise} total).

Dataset	Noise	p_{noise}	Open:Real	Open:Noise
GBSG ($p = 7$)	0×	0	3.89 ± 0.87	0.00 ± 0.00
	3×	21	3.56 ± 0.68	0.00 ± 0.00
	5×	35	3.22 ± 0.79	0.00 ± 0.00
	7×	49	3.56 ± 0.83	0.00 ± 0.00
	10×	70	3.11 ± 0.74	0.00 ± 0.00
METABRIC ($p = 9$)	0×	0	4.56 ± 0.83	0.00 ± 0.00
	3×	27	4.89 ± 1.37	0.11 ± 0.31
	5×	45	4.67 ± 0.67	0.33 ± 0.67
	7×	63	4.78 ± 0.79	0.22 ± 0.42
	10×	90	4.00 ± 0.67	0.00 ± 0.00
SUPPORT ($p = 14$)	0×	0	5.56 ± 0.50	0.00 ± 0.00
	3×	42	5.22 ± 0.63	0.00 ± 0.00
	5×	70	5.11 ± 0.57	0.00 ± 0.00
	7×	98	5.11 ± 0.57	0.00 ± 0.00
	10×	140	5.11 ± 0.74	0.00 ± 0.00
AIDS ($p = 11$)	0×	0	3.33 ± 1.63	0.00 ± 0.00
	3×	33	2.22 ± 1.31	2.89 ± 1.45
	5×	55	1.56 ± 0.96	3.89 ± 1.66
	7×	77	1.22 ± 0.92	3.22 ± 1.03
	10×	110	0.11 ± 0.31	4.67 ± 1.83
FLCHAIN ($p = 8$)	0×	0	3.00 ± 0.00	0.00 ± 0.00
	3×	24	3.00 ± 0.00	0.00 ± 0.00
	5×	40	3.00 ± 0.00	0.00 ± 0.00
	7×	56	3.00 ± 0.00	0.00 ± 0.00
	10×	80	3.00 ± 0.00	0.00 ± 0.00
NWTCO ($p = 3$)	0×	0	2.11 ± 0.57	0.00 ± 0.00
	3×	9	2.67 ± 0.47	0.22 ± 0.42
	5×	15	2.67 ± 0.67	0.22 ± 0.42
	7×	21	2.56 ± 0.50	0.22 ± 0.42
	10×	30	2.67 ± 0.47	0.22 ± 0.63

SUPPORT showing exact 0.00±0.00 noise-gate activation throughout. Across high-censoring datasets (AIDS, FLCHAIN, NWTCO), FLCHAIN and NWTCO show near or fully-complete suppression of injected noise features. AIDS is the clear exception: noise-gate leakage grows with noise multiplier, reaching 4.67 ± 1.83 open noise gates at 10×, while the number of open *real* feature gates collapses toward zero over the same range (from 3.33 at 0× to 0.11 at 10×). This pattern is consistent with AIDS's extreme censoring rate (92%): with very few observed events, the local KM neighborhoods used for imputation (Section 3.3) provide a weak and noisy training signal, making it difficult for the sparsity-inducing ℓ_0 penalty to reliably distinguish informative from spurious features. This reinforces the boundary

condition noted in Section 5.1: under extreme censoring, GRAFT's advantages including reliable feature selection diminish, and simpler parametric baselines may be preferable.

C.8. D-Calibration Results

In addition to the Integrated Brier Score reported in the main text, we assess calibration using D-Calibration (Haider et al., 2020). For an uncensored subject, evaluating the model's survival function at that subject's observed event time yields a value that should be uniformly distributed on $[0, 1]$ under a well-calibrated model, by the probability integral transform. Censored subjects are incorporated via Haider et al. (2020) censoring-adjusted procedure, which distributes each censored subject's contribution across the bins consistent with their (unobserved) true event time, such that the resulting combined statistic is uniformly distributed under the true model. In practice, these values are partitioned into $B = 10$ bins and compared against the expected uniform proportion using a Pearson χ^2 goodness-of-fit test with 9 degrees of freedom. A model is considered D-calibrated at a given seed if $p > 0.05$. GRAFT's survival curves are obtained using the same post-hoc CoxPH calibration procedure described in Section 3.6.

Table 10 reports per-seed D-Calibration p-values across all six benchmark datasets and all models, using the same 3-fold, 3-seed evaluation protocol as the main experiments, along with the number of seeds (out of 3) for which each model was D-calibrated. On AIDS, all models except DeepHit achieve near-perfect p-values close to 1.0; this is likely a consequence of the dataset's extreme censoring rate (92%), under which the censoring-adjustment term in the D-Calibration statistic spreads probability mass across many bins for most subjects, making the test less sensitive to genuine miscalibration at this censoring level.

Overall, GRAFT is D-calibrated on all datasets and seeds achieving the highest p values followed by CoxPH and RSF which are both D-calibrated across all seeds on 5/6 datasets. DeepHit fails to achieve D-Calibration on any dataset, and DeepSurv and Weibull are D-calibrated only intermittently. SUPPORT is the one dataset where D-Calibration is difficult to achieve across all models considered. SUPPORT is both the largest of the six benchmarks (8,873 subjects) and enrolls a clinically heterogeneous population of critically ill patients across nine distinct disease categories (*e.g.*, acute respiratory failure, congestive heart failure, coma, and several cancer types) (Knaus et al., 1995), with survival times ranging from days to several years and a median follow-up of only 231 days. Prognosis is known to vary substantially across these disease groups (Lynn et al., 1997), and this population heterogeneity likely makes it inherently more difficult for a covariate-based model to produce individually well-calibrated curves across the full range of outcomes.

Table 10: D-Calibration results (Haider et al., 2020) across all six datasets and all models, based on 3-fold cross-validation with 3 random seeds. p-values are reported individually per seed; *Passed* indicates the number of seeds (out of 3) with $p > 0.05$. Higher p-values and more seeds passed indicate better calibration.

Dataset	**Model**	p **(Seed 1)**	p **(Seed 2)**	p **(Seed 3)**	**Passed / 3**
Low Censoring Datasets					
GBSG	GRAFT	0.9234	0.8655	0.8414	3
	CoxPH	0.2824	0.5526	0.1352	3
	Weibull	0.0000	0.0000	0.0000	0
	DeepHit	0.0000	0.0000	0.0000	0
	DeepSurv	0.0000	0.0000	0.0000	0
	RSF	0.5758	0.6038	0.4426	3
METABRIC	GRAFT	0.9933	0.6891	0.9782	3
	CoxPH	0.7691	0.6349	0.3826	3
	Weibull	0.0961	0.1081	0.1009	3
	DeepHit	0.0001	0.0001	0.0000	0
	DeepSurv	0.4496	0.2226	0.6417	3
	RSF	0.9902	0.8894	0.4187	3
SUPPORT	GRAFT	0.1124	0.0831	0.0787	3
	CoxPH	0.0966	0.0193	0.0253	1
	Weibull	0.0000	0.0000	0.0000	0
	DeepHit	0.0000	0.0000	0.0000	0
	DeepSurv	0.0000	0.0000	0.0000	0
	RSF	0.0000	0.0000	0.0000	0
High Censoring Datasets					
AIDS	GRAFT	1.0000	1.0000	1.0000	3
	CoxPH	1.0000	1.0000	1.0000	3
	Weibull	1.0000	1.0000	1.0000	3
	DeepHit	0.0000	0.0000	0.0000	0
	DeepSurv	0.9995	0.9990	0.9999	3
	RSF	1.0000	1.0000	1.0000	3
FLCHAIN	GRAFT	0.9275	0.9336	0.9288	3
	CoxPH	0.8855	0.9484	0.7354	3
	Weibull	0.3573	0.3237	0.2883	3
	DeepHit	0.0000	0.0000	0.0000	0
	DeepSurv	0.0001	0.0000	0.0000	0
	RSF	0.7146	0.7826	0.8400	3
NWTCO	GRAFT	0.9959	0.9985	0.9968	3
	CoxPH	1.0000	1.0000	0.9999	3
	Weibull	0.4276	0.4240	0.3240	3
	DeepHit	0.0000	0.0000	0.0000	0
	DeepSurv	0.0572	0.0150	0.0000	1
	RSF	0.9983	0.9992	1.0000	3